Data Modeling in Business (BMGT 430)

Project Title: **Life Expectancy Analysis**
Group Members: **Maria Master, Ayda Hailemariam, Tina Zhang, Hong Anh Nguyen, Elise Manglicmot**

I pledge my honor that I have not given or received any unauthorized assistance on this project

---

Team member	**Signature**
Maria Gomes Master	
Ayda Hailemariam	*Ayda Hailemariam*
Tina Zhang	*Tina Zhang*
Hong Anh Nguyen	*Hong Anh Nguyen*
Elise Manglicmot	*Elise Manglicmot*

Life Expectancy Analysis

LIFE Expectancy Data Description

We found our selected dataset on average life expectancy in various countries and the potential influencing factors on Kaggle. It has been compiled using data from the World Health Organization and the World Bank. After narrowing it down to the specific year 2015 (the most recent year in the data), our data set has a sample size of n = 179 countries and k = 14 predictors.

We are interested in this dataset on average life expectancy and its factors for a variety of reasons. Understanding these factors is critical for developing focused public health interventions. We may personalize community health policies by looking at characteristics such as infant mortality, illness prevalence, lifestyle choices, and socioeconomic determinants. Analyzing regional differences in life expectancy can help influence resource allocation and policymaking to address health inequities.

Our Response Variable is:

Response	**Description**	**Units**	**Category**
LIFEEXP	The average life expectancy of both genders for a country (since birth)	years	Numerical

Our **Numerical Predictors** for our response variable, LIFFEXP, consist of:

Predictors	**Description** (per country)	**Units**
INFDTH	the number of infant deaths	deaths per 1000 population
FIVEDTH	the number of children (< 5 years) deaths	deaths per 1000 population
ADTMORT	the number of adult deaths	deaths per 1000 population
ALCOHOL	the alcohol consumption of people 15+ years old	liters of pure alcohol per capita
HBV	coverage of Hepatitis B (HepB3) immunization among 1-year-olds	% of 1-year-olds population
MMR	coverage of Measles-containing vaccine first dose (MCV1) immunization among 1-year-olds	% of 1-year-olds population
BMI	the average BMI for that country	kg/m^2
POLIO	coverage of Polio (Pol3) immunization among 1-year-olds	% of 1-year-olds population
DIPH	coverage of Diphtheria tetanus toxoid and pertussis (DTP3) immunization among 1-year-olds.	% of 1-year-olds population

HIVINC	number of incidents of HIV for people aged 15-49	incidents per 1000 population
GDP	GDP per capita of that country	United States Dollars
POP	total population	millions of people
EDUC	average years that people aged 25+ spent in formal education	years

Our **Categorical Predictors** for our response variable, LIFEEXP, consists of:

Predictors	**Description**	**Baseline (left out of model)**	**Other Indicators**
ECONOMY	the economic conditions of that country	Developed	Developing
REGION	the region of the country was distributed into	AFR (Africa)	-ASIA -ME (Middle East) -EU (European Nation) -REURO (Rest of Europe) -SA (South America) -NA (North America) -OCEANIA -CAC (Central America and the Caribbean)

Research Question

Our project aims to construct a model that can predict the average life expectancy of a country. To build this model, the overarching question we are trying to answer is "What predictors significantly affect life expectancy?" To answer this question, we have to determine which of the predictors are insignificant to the model and if these can be removed. Afterwards, we will try to answer questions such as:

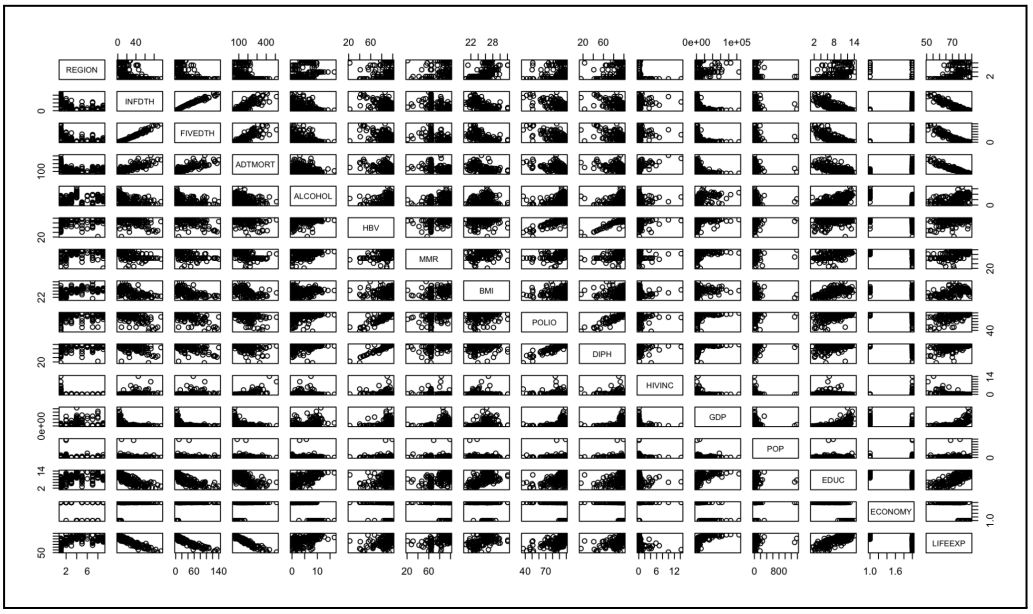
1) Which factors influence life expectancy the most?
2) Which region has the greatest effect on increasing life expectancy?

Model Building

After we organized our data set, we coded our categorical predictors with indicator variables in R. Then, we ran an initial regression analysis to see what kind of data we were working with. As we analyzed the regression model, we made sure to keep some questions in mind:

1) Is the model significant?
2) Does the model fulfill all four LINE Assumptions?
3) Are there any predictors with a p-value greater than $\alpha = 0.05$ or 5%?
4) Can any (or all) of these insignificant predictors be removed from the model?

Graphing Scatterplot Matrix for the initial data set:



Summary Statistics for the data set:

```
    REGION       INFDTH         FIVEDTH          ADTMORT          ALCOHOL
AFR    :51   Min.   : 1.80   Min.   :  2.30   Min.   : 49.38   Min.   : 0.000
ASIA   :27   1st Qu.: 6.65   1st Qu.:  7.85   1st Qu.: 90.79   1st Qu.: 1.360
EU     :27   Median :15.20   Median : 17.50   Median :146.52   Median : 4.040
CAC    :19   Mean   :23.56   Mean   : 31.68   Mean   :163.67   Mean   : 4.729
REURO  :15   3rd Qu.:36.55   3rd Qu.: 49.95   3rd Qu.:215.65   3rd Qu.: 7.760
ME     :14   Max.   :95.10   Max.   :140.20   Max.   :513.48   Max.   :16.720
(Other):26
     HBV              MMR              BMI             POLIO            DIPH
Min.   :22.0   Min.   :21.00   Min.   :20.5   Min.   :37.00   Min.   :16.00
1st Qu.:82.5   1st Qu.:64.00   1st Qu.:23.8   1st Qu.:85.00   1st Qu.:85.50
Median :92.0   Median :84.00   Median :26.2   Median :93.00   Median :93.00
Mean   :87.1   Mean   :80.23   Mean   :25.6   Mean   :88.26   Mean   :87.92
3rd Qu.:97.0   3rd Qu.:94.00   3rd Qu.:27.0   3rd Qu.:97.00   3rd Qu.:97.00
Max.   :99.0   Max.   :99.00   Max.   :32.1   Max.   :99.00   Max.   :99.00

    HIVINC             GDP              POP               EDUC             ECONOMY
Min.   : 0.0100   Min.   :   306   Min.   :   0.090   Min.   : 1.400   Developed : 37
1st Qu.: 0.0800   1st Qu.:  1690   1st Qu.:   2.215   1st Qu.: 5.950   Developing:142
Median : 0.1400   Median :  5391   Median :   9.110   Median : 8.700
Mean   : 0.6098   Mean   : 12617   Mean   :  40.088   Mean   : 8.361
3rd Qu.: 0.3700   3rd Qu.: 14274   3rd Qu.:  27.445   3rd Qu.:11.050
Max.   :14.3000   Max.   :105462   Max.   :1379.860   Max.   :14.100

   LIFEEXP
Min.   :50.90
1st Qu.:66.30
Median :73.00
Mean   :71.46
3rd Qu.:76.85
Max.   :83.80
```

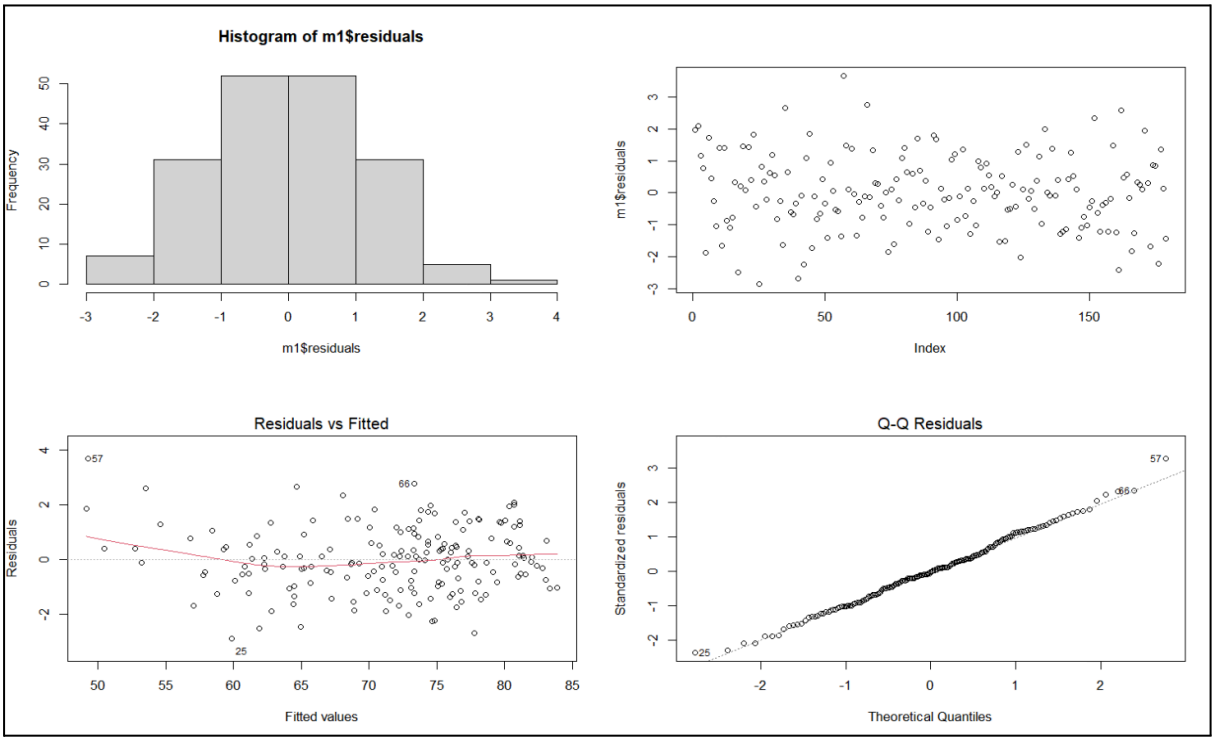
Initial Regression Output created in R Studio:

```
Call:
lm(formula = LIFEEXP ~ ., data = dat2)

Residuals:
    Min      1Q  Median      3Q     Max
-2.8702 -0.7814 -0.0348  0.7334  3.6714

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)         8.731e+01  2.201e+00  39.672  < 2e-16 ***
REGIONASIA          2.386e-01  4.111e-01   0.580 0.562551
REGIONCAC           1.878e+00  4.774e-01   3.934 0.000125 ***
REGIONEU           -6.487e-01  6.801e-01  -0.954 0.341641
REGIONME            1.428e-01  5.206e-01   0.274 0.784247
REGIONNA            1.033e-02  9.454e-01   0.011 0.991293
REGIONOCEANIA      -8.718e-01  5.596e-01  -1.558 0.121284
REGIONREURO         2.131e-01  5.452e-01   0.391 0.696429
REGIONSA            1.930e+00  5.215e-01   3.701 0.000297 ***
INFDTH             -2.838e-02  3.781e-02  -0.750 0.454140
FIVEDTH            -6.722e-02  2.595e-02  -2.591 0.010486 *
ADTMORT            -4.997e-02  2.970e-03 -16.824  < 2e-16 ***
ALCOHOL            -1.219e-02  4.902e-02  -0.249 0.803952
HBV                -1.657e-02  2.402e-02  -0.690 0.491361
MMR                 1.095e-02  7.974e-03   1.373 0.171698
BMI                -1.605e-01  7.642e-02  -2.101 0.037292 *
POLIO              -1.609e-03  2.223e-02  -0.072 0.942380
DIPH                7.822e-03  2.697e-02   0.290 0.772160
HIVINC              2.098e-01  8.996e-02   2.333 0.020951 *
GDP                 2.485e-05  8.866e-06   2.803 0.005700 **
POP                -2.425e-04  6.893e-04  -0.352 0.725503
EDUC                8.966e-02  7.293e-02   1.229 0.220749
ECONOMYDeveloping -2.657e+00  6.351e-01  -4.184 4.77e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.244 on 156 degrees of freedom
Multiple R-squared:  0.9779,    Adjusted R-squared:  0.9748
F-statistic: 313.4 on 22 and 156 DF,  p-value: < 2.2e-16
```

Our Initial Model:

LIFEEXP = β_0+ β_1REGION_ASIA + β_2REGION_CAC - β_3REGION_EU + β_4REGION_ME + β_5REGION_NA - β_6REGION_OCEANIA + β_7REGION_REURO + β_8REGION_SA - β_9INFDTH - β_{10}FIVEDTH - β_{11}ADTMORT - β_{12}ALCOHOL - β_{13}HBV + β_{14}MMR - β_{15}BMI - β_{16}POLIO + β_{17}DIPH + β_{18}HIVINC+ β_{19}GDP - β_{20}POP + β_{21}EDU - β_{22}ECONOMY_ developing + e

Graphs to Analyze the LINE Assumptions of the Original Model:



Assumption	Graph	Decision	Reasoning
Linearity	Residual vs Fitted	Satisfied	The mean of the points within the slices seems to be roughly equal to 0
Independence	Residuals vs Index	Satisfied	There is no pattern in the spread of the points.
Normality	Histogram of residuals Q-Q plot	Satisfied	The histogram of residuals looks roughly normally distributed Most of the points in the Q-Q plot do not significantly deviate from the line with a slope of 45 degrees
Equal Variance	Residual vs Fitted	Satisfied	The spread of the points within the slices is roughly the same.

Reduced Model

Despite the original model fulfilling all four assumptions, we noticed in the regression output that several predictors had very high p-values. In an effort to create a parsimonious model, we conducted a **T-test for each predictor** and indicator variables to determine if that predictor was significant for the model. We summarized our results in the table below:

Predictor (abbrev)	**T-test (α = 5%)**
Hypotheses for all the T-tests:	Hypothesis: - Ho: $\beta_i = 0$ - Ho: $\beta_i \neq 0$
Test statistic for all T-tests	T obs/ T value from regression output
β_2 REGION_CAC β_8 REGION_SA β_{10} FIVEDTH β_{11} ADTMORT β_{15} BMI β_{18} HIVINC β_{19} GDP β_{22}ECONOMY_ developing *β_1REGION_ASIA* *β_3REGION_EU* *β_4REGION_ME* *β_5REGION_NA* *β_6REGION_OCEANIA* *β_7REGIONREURO*	Decision Rule: Since all of these predictor's p-value < alpha 0.05, we reject Ho **Conclusion**: These predictors (to the left) are significant predictors in explaining the response while holding the other predictors fixed. **While some of the Region dummy variables were deemed insignificant predictors according to the T-test, we decided to keep them all in our model because two of them were significant. Since even one was significant, we believe that in this context, the region of a country could play a major part in predicting life expectancy.*
all the other β's (8 in total)	Decision Rule: Since all of these predictor's p-value > alpha 0.05, we fail to reject Ho **Conclusion**: These 8 predictors are each not a significant predictor in explaining the response after adjusting for the other predictors

We will now conduct a **Partial F-test** to determine if all these other β's can be simultaneously removed from the model at the same time:

Models	**Model$_{FULL}$:** β_0+ β_1REGION_ASIA + β_2 REGION_CAC - β_3REGION_EU + β_4REGION_ME + β_5REGION_NA - β_6REGION_OCEANIA + β_7REGION_REURO + β_8REGION_SA - β_9 INFDTH - β_{10} FIVEDTH - β_{11} ADTMORT - β_{12} ALCOHOL - β_{13} HBV + β_{14} MMR - β_{15} BMI - β_{16} POLIO + β_{17}DIPH + β_{18} HIVINC+ β_{19} GDP - β_{20}POP + β_{21}EDUC - β_{22}ECONOMY_ developing + e **Model$_{REDUCED}$**: β_0+ β_1REGION_ASIA + β_2 REGION_CAC - β_3REGION_EU + β_4REGION_ME + β_5REGION_NA - β_6REGION_OCEANIA + β_7REGION_REURO + β_8REGION_SA - β_{10}FIVEDTH - β_{11} ADTMORT - β_{15}BMI + β_{18} HIVINC+ β_{19} GDP - β_{22}ECONOMY_ developing + e

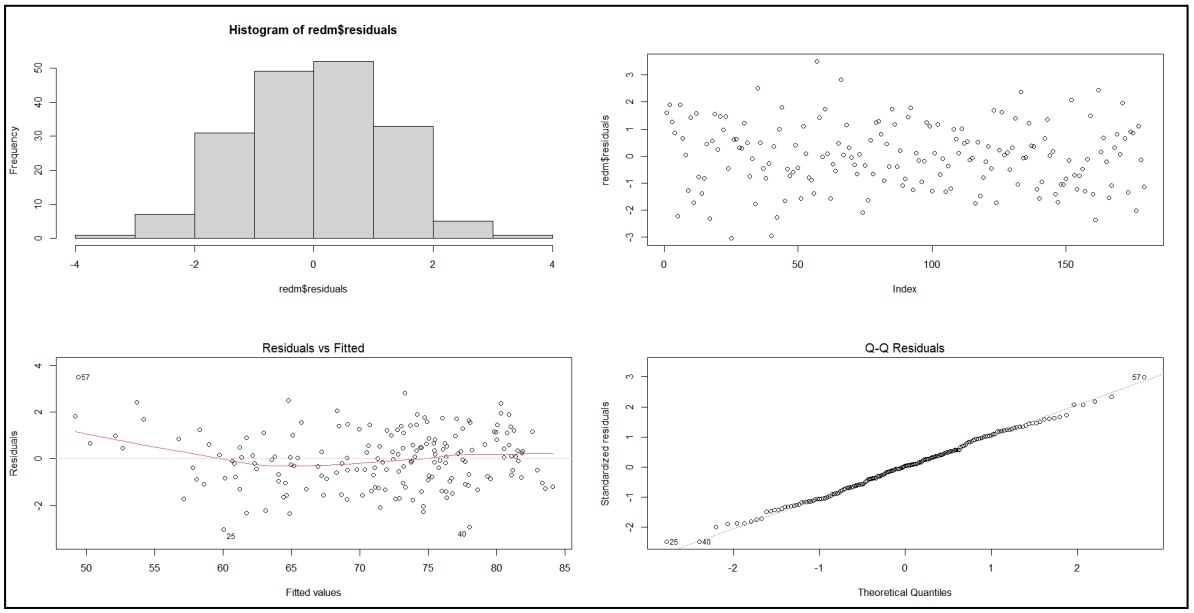
ANOVA TABLE	Model 1: LIFEEXP ~ REGION + FIVEDTH + ADTMORT + BMI + HIVINC + GDP + ECONOMY Model 2: LIFEEXP ~ REGION + INFDTH + FIVEDTH + ADTMORT + ALCOHOL + HBV + MMR + BMI + POLIO + DIPH + HIVINC + GDP + POP + EDUC + ECONOMY Res.Df RSS Df Sum of Sq F Pr(>F) 1 164 250.30 2 156 241.56 8 8.7394 0.7055 0.6864
Step 1: Hypothesis	Ho: $\beta 9 = \beta 12 = \beta 13 = \beta 14 = \beta 16 = \beta 17 = \beta 20 = \beta 21 = 0$ Ha: at least one of the β's $\neq 0$
Step 2: Test Statistic	F (8:156) = 0.7055 (from ANOVA table) OR Formula → $F = (SSE_R\text{-}SSE_F)/(df_R\text{-}df_F) \div SSE_F/df_F$ F (8:156)= (250.30-241.56)/8÷241.56/156 = 0.7055
Step 3: P-value	P-value = 0.6864 (from ANOVA table)
Step 4: Compare p-value and alpha	P-value: 0.6864 > alpha: 0.05 Decision Rule: p-value > alpha → fail to reject Ho **<u>Conclusion</u>:** Since the p-value, 0.6864 > alpha 0.05, we failed to reject Ho, hence we can eliminate all 8 predictors from the model at the same time. **Therefore, we can use the reduced model**.
Regression output for the Reduced Model	**Regression Output** for the **Reduced Model (redM$_1$)** created in R Studio: Residuals: Min 1Q Median 3Q Max -3.0421 -0.8353 0.0300 0.8325 3.5073 Coefficients: Estimate Std. Error t value Pr(>|t|) (Intercept) 8.642e+01 1.810e+00 47.752 < 2e-16 *** REGIONASIA 2.942e-01 3.672e-01 0.801 0.424218 REGIONCAC 1.881e+00 4.345e-01 4.330 2.59e-05 *** REGIONEU -5.857e-01 6.420e-01 -0.912 0.362959 REGIONME 1.678e-01 4.902e-01 0.342 0.732559 REGIONNA 2.307e-01 8.605e-01 0.268 0.788937 REGIONOCEANIA -9.615e-01 5.143e-01 -1.870 0.063302 . REGIONREURO 5.403e-01 4.871e-01 1.109 0.268967 REGIONSA 1.863e+00 4.848e-01 3.844 0.000173 *** FIVEDTH -8.716e-02 7.109e-03 -12.261 < 2e-16 *** ADTMORT -5.001e-02 2.814e-03 -17.771 < 2e-16 *** BMI -9.589e-02 6.668e-02 -1.438 0.152322 HIVINC 1.966e-01 8.586e-02 2.290 0.023276 * GDP 2.818e-05 8.512e-06 3.310 0.001146 ** ECONOMYDeveloping -2.916e+00 5.897e-01 -4.944 1.87e-06 *** --- Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

The **Reduced Model (redM$_1$)**:

LIFEEXP: β_0+ β_1REGION_ASIA + β_2 REGION_CAC - β_3REGION_EU + β_4REGION_ME + β_5REGION_NA - β_6REGION_OCEANIA + β_7REGION_REURO + β_8REGION_SA - β_9FIVEDTH - β_{10} ADTMORT - β_{11} BMI + β_{12} HIVINC + β_{13} GDP + β_{14}ECONOMY_ developing + e

Looking at the **adjusted R-squared** values for the reduced model and the original model, we can see that the value increased after removing the 8 predictors. The adjusted R-squared value increased from .9748 to .**9751**, indicating that we did remove predictors that were insignificant as the higher the adjusted R-squared value is, the better the model is.

Now, we will be analyzing the **LINE Assumptions** for this **Reduced Model (redM$_1$)** to see if it satisfies the assumptions better than the initial model:



Assumption	Graph	Decision	Reasoning
Linearity	Residual vs Fitted	Satisfied	The mean of the points within the slices seems to be roughly equal to 0
Independence	Residuals vs Index	Satisfied	There is no pattern in the spread of the points.
Normality	Histogram of residuals Q-Q plot	Satisfied (better than the original model)	The histogram of residuals looks roughly normally distributed, more so than the original histogram Most of the points in the Q-Q plot do not significantly deviate from the line with a slope of 45 degrees, more so than the original Q-Q plot
Equal Variance	Residual vs Fitted	Satisfied (better than the original model)	The spread of the points within the slices is roughly the same, more so than the original Residual vs Fitted plot

While this reduced model does fulfill the LINE assumptions slightly better than the original model did (in terms of normality and equal variance), we wanted to apply transformations to this cleaned-up, reduced model to see if we could make the model even better.

Transformations

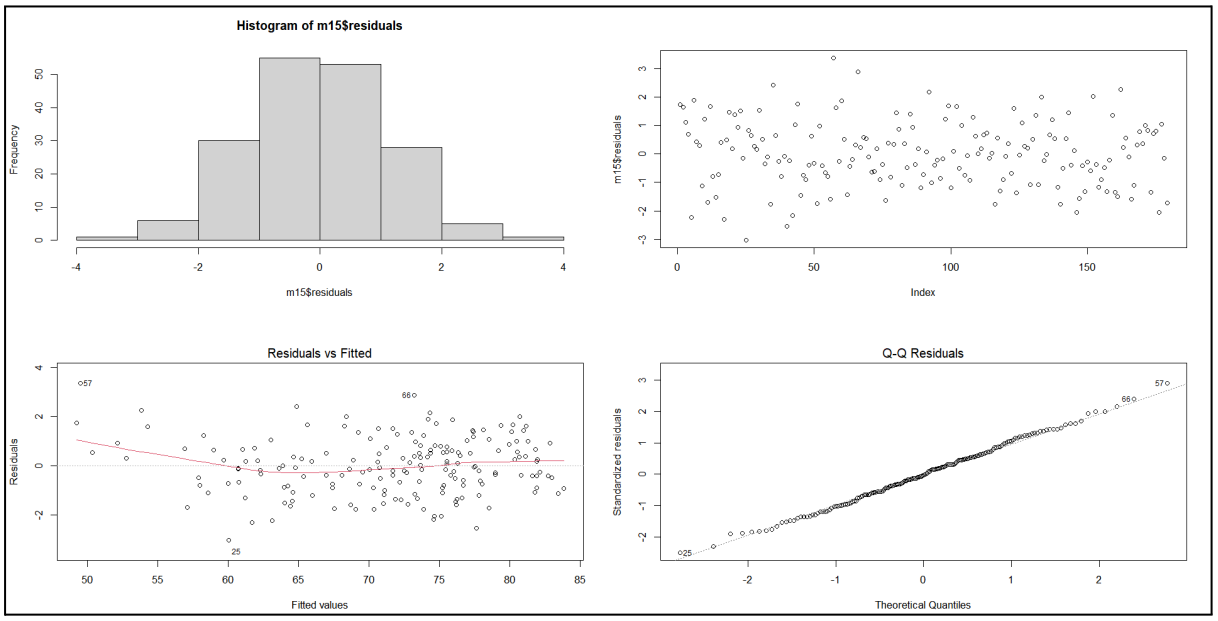
Model 1: Transformation 1: Interaction Terms

Firstly, to improve the reduced model, we added interaction terms such as ECONOMY:GDP where ECONOMY is a categorical variable and GDP is a numerical variable, and REGION:BMI where REGION is a categorical variable and BMI is a numerical variable. ECONOMY:GDP interaction term considers how the impact of GDP (a measure of economic prosperity) on life expectancy might differ between developed and developing economies. REGION:BMI interaction term examines how the association between BMI (Body Mass Index) and life expectancy varies across different regions.

Regression Output for the **Transformed Model with Interaction Terms**:

```
Call:
lm(formula = LIFEEXP ~ REGION + FIVEDTH + ADTMORT + BMI + HIVINC +
    GDP + ECONOMY + REGION:BMI + ECONOMY:GDP, data = dat2)

Residuals:
    Min      1Q  Median      3Q     Max
-3.0320 -0.7892 -0.0436  0.7305  3.3705

Coefficients:
                       Estimate Std. Error t value Pr(>|t|)
(Intercept)           8.460e+01  2.981e+00  28.376  < 2e-16 ***
REGIONASIA            5.080e+00  4.210e+00   1.206   0.2295
REGIONCAC            -9.449e+00  7.126e+00  -1.326   0.1868
REGIONEU              2.144e+01  1.071e+01   2.003   0.0470 *
REGIONME              8.236e+00  6.225e+00   1.323   0.1878
REGIONNA              1.858e+01  2.578e+01   0.721   0.4723
REGIONOCEANIA         8.282e-01  4.350e+00   0.190   0.8493
REGIONREURO           2.287e+00  1.605e+01   0.143   0.8869
REGIONSA             -9.509e+00  1.807e+01  -0.526   0.5996
FIVEDTH              -8.593e-02  7.181e-03 -11.967  < 2e-16 ***
ADTMORT              -4.905e-02  2.861e-03 -17.143  < 2e-16 ***
BMI                  -8.780e-03  1.149e-01  -0.076   0.9392
HIVINC                1.607e-01  8.965e-02   1.792   0.0750 .
GDP                   1.987e-05  1.118e-05   1.778   0.0774 .
ECONOMYDeveloping    -3.466e+00  7.397e-01  -4.686 6.06e-06 ***
REGIONASIA:BMI       -1.972e-01  1.738e-01  -1.135   0.2583
REGIONCAC:BMI         4.105e-01  2.677e-01   1.533   0.1273
REGIONEU:BMI         -8.479e-01  4.103e-01  -2.066   0.0404 *
REGIONME:BMI         -3.048e-01  2.329e-01  -1.309   0.1925
REGIONNA:BMI         -6.681e-01  9.263e-01  -0.721   0.4718
REGIONOCEANIA:BMI    -7.402e-02  1.653e-01  -0.448   0.6549
REGIONREURO:BMI      -6.940e-02  6.071e-01  -0.114   0.9091
REGIONSA:BMI          4.171e-01  6.762e-01   0.617   0.5383
GDP:ECONOMYDeveloping 1.745e-05  1.791e-05   0.974   0.3314
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.225 on 155 degrees of freedom
Multiple R-squared:  0.9787,     Adjusted R-squared:  0.9755
F-statistic: 309.4 on 23 and 155 DF,  p-value: < 2.2e-16
```

Analyze the LINE Assumptions of the Transformed Model with Interaction Terms:



Assumption	**Graph**	**Decision**	**Reasoning**
Linearity	Residual vs Fitted	Satisfied	The mean of the points within the slices seems to be roughly equal 0
Independence	Residuals vs Index	Satisfied	There is no pattern in the spread of the points
Normality	Histogram of residuals Q-Q plot	Satisfied	The histogram of residuals looks roughly normally distributed Most of the points in the Q-Q plot do not significantly deviate from the line with a slope of 45 degrees
Equal Variance	Residual vs Fitted	Satisfied	The spread of the points within the slices is roughly the same, more so than the original Residual vs Fitted plot

Is the transformed interaction terms model significant?

To answer this question, we performed a **Global F-test** with $\alpha = 5\%$:

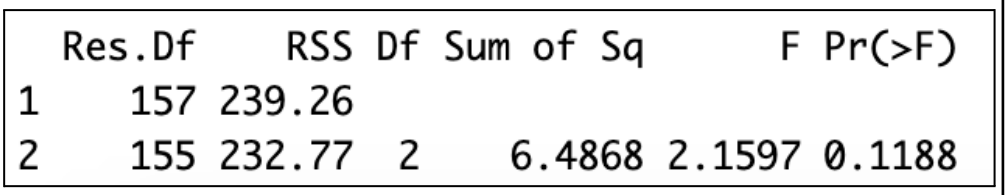
Step 1: Hypothesis	Ho: $\beta 1 = \beta 2 = \ldots \beta 23 = 0$ Ha: at least one of the β's $\neq 0$
Step 2: Test Statistic	F(23:155) = 309.4, where k=23 and n=179
Step 3: P-value	P < 2.2e-16 → P = approx. 0
Step 4:	P-value: approx. 0 < alpha: 0.05

Compare p-value and alpha	Decision Rule: p-value < alpha → reject Ho **<u>Conclusion:</u>** Since our p-value, 0, is less than alpha, 0.05, we reject our null hypothesis. At least one of the predictors is significant in explaining the response, therefore the **transformed model with interaction terms is a significant model.**

Are there any Insignificant Predictors in this Transformed Interaction Terms Model?
To answer this question, we conducted **T-tests for each of the predictors.** As we had 23 predictors, the table below is a summary of our T-test results:

Predictor	**T-test (α = 5%)**
Hypotheses for all the T-tests:	Hypothesis: - Ho: $\beta_i = 0$ - Ho: $\beta_i \neq 0$
β_3REGION_EU β_9FIVEDTH β_{10}ADTMORT β_{13}GDP β_{14}Economy_Developing β_{17}REGION_EU:BMI *β_1REGION_ASIA** *β_2REGION_CAC* *β_4REGION_ME* *β_5REGION_NA* *β_6REGION_OC* *β_7REGIONREURO* *β_8REGION_SA* *β_{11}BMI*** *β_{15}REGION_ASIA:BMI**** *β_{16} REGION_CAC:BMI* *β_{18}REGION_ME:BMI* *β_{19}REGION_NA:BMI* *β_{20}REGION_OC:BMI* *β_{21}REGIONREURO:BMI* *β_{22}REGION_SA:BMI*	Decision Rule: Since all of these predictor's p-value < alpha 0.05, we reject Ho **Conclusion**: These predictors (to the left) are each significant predictors in explaining the response after adjusting for the other predictors **While some of the Region dummy variables were deemed insignificant predictors according to the T-test, we decided to keep them in our model because one of them was significant.* ***BMI itself isn't significant. but due to the Hierarchy of Principle of interaction terms, we kept it in the model.* ****While some of the interaction terms were deemed as insignificant predictors according to the T-test, we decided to keep them in our model because one of them was significant.*
all the other β's (2 in total)	Decision Rule: Since all of these predictor's p-value > alpha 0.05, we fail to reject Ho **Conclusion**: These 2 predictors are each not a significant predictor in explaining the response after adjusting for the other predictors

We will now conduct a **Partial F-test** to determine if all these other β's can be simultaneously removed from the model at the same time:

Models	**Model**$_{FULL}$: β_0+ β_1REGION_ASIA + β_2 REGION_CAC - β_3REGION_EU + β_4REGION_ME + β_5REGION_NA - β_6REGION_OCEANIA + β_7REGION_REURO + β_8REGION_SA - β_9 FIVEDTH - β_{10} ADTMORT - β_{11} BMI + β_{12} HIVINC+ β_{13} GDP + β_{14}ECONOMY_ developing - β_{15}REGION_ASIA:BMI +β_{16} REGION_CAC:BMI - β_{17} REGION_EU:BMI - β_{18}REGION_ME:BMI - β_{19}REGION_NA:BMI - β_{20}REGION_OCEANIA:BMI - β_{21}REGIONREURO:BMI + β_{22}REGION_SA:BMI + β_{23}GDP:ECONOMYDeveloping + e **Model**$_{REDUCED}$: β_0+ β_1REGION_ASIA + β_2 REGION_CAC - β_3REGION_EU + β_4REGION_ME + β_5REGION_NA - β_6REGION_OCEANIA + β_7REGION_REURO + β_8REGION_SA - β_9 FIVEDTH - β_{10} ADTMORT - β_{11} BMI + β_{13}GDP + β_{14}ECONOMY_ developing - β_{15}REGION_ASIA:BMI +β_{16}REGION_CAC:BMI - β_{17}REGION_EU:BMI - β_{18}REGION_ME:BMI - β_{19}REGION_NA:BMI - β_{20}REGION_OCEANIA:BMI - β_{21}REGIONREURO:BMI + β_{22}REGION_SA:BMI + e
Step 1: Hypothesis	Ho: $\beta 12 = \beta 23 = 0$ Ha: at least one of the β's $\neq 0$
Step 2: Test Statistic	F (2:155) = 2.1597 (from ANOVA table in R)
Step 3: P-value	P-value = 0.1188 (from ANOVA table in R)
Step 4: Compare p-value and alpha	P-value: > alpha: 0.05 Decision Rule: **p-value > alpha → fail to reject Ho** (hence use reduced model)
ANOVA TABLE	

	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
1	157	239.26				
2	155	232.77	2	6.4868	2.1597	0.1188

Since the p-value, 0.1188 > alpha 0.05, we failed to reject Ho, hence we can eliminate the 2 predictors from the model at the same time. We can use the **Reduced Model (redM$_2$)**:

LIFEEXP: β_0+ β_1REGION_ASIA + β_2 REGION_CAC - β_3REGION_EU + β_4REGION_ME + β_5REGION_NA - β_6REGION_OCEANIA + β_7REGION_REURO + β_8REGION_SA - β_9 FIVEDTH - β_{10} ADTMORT - β_{11} BMI + β_{12}GDP - β_{13}ECONOMY_ developing - β_{14}REGION_ASIA:BMI +β_{15}REGION_CAC:BMI - β_{16}REGION_EU:BMI - β_{17}REGION_ME:BMI - β_{18}REGION_NA:BMI - β_{19}REGION_OCEANIA:BMI - β_{20}REGIONREURO:BMI + β_{21}REGION_SA:BMI + e

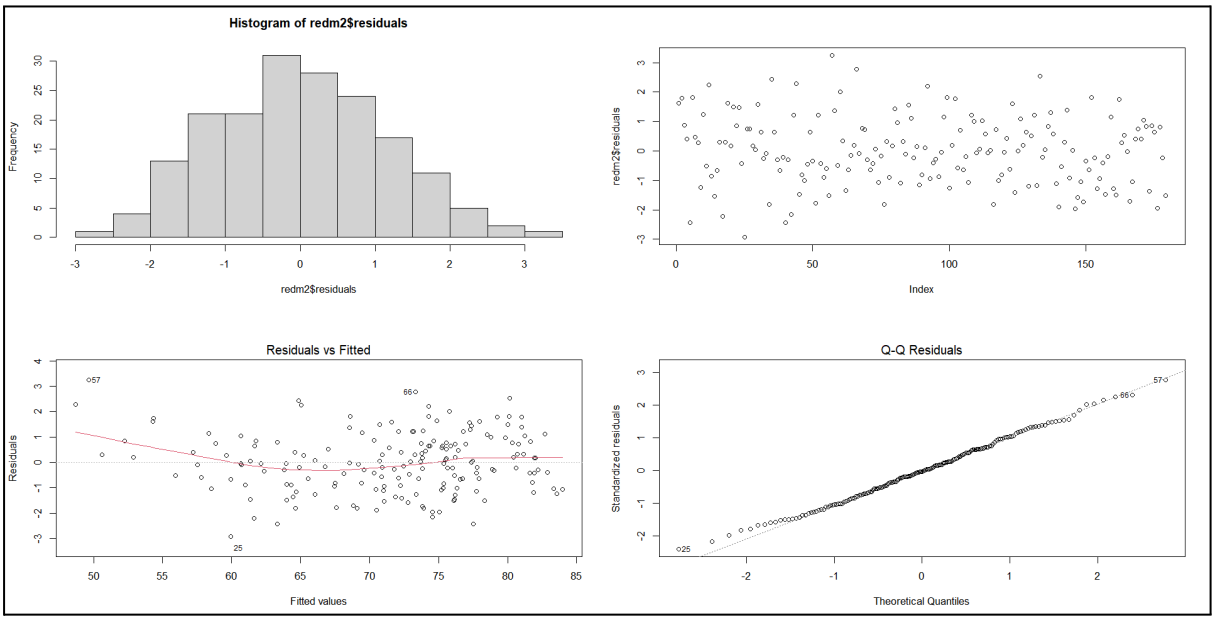
Regression Output for the **Reduced Model with Interaction Terms (redM$_2$):**

```
Call:
lm(formula = LIFEEXP ~ REGION + FIVEDTH + ADTMORT + BMI + GDP +
    ECONOMY + REGION:BMI, data = dat1)

Residuals:
    Min      1Q  Median      3Q     Max
-2.9356 -0.8649 -0.0422  0.7622  3.2534

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)        8.248e+01  2.808e+00  29.369  < 2e-16 ***
REGIONASIA         6.612e+00  4.114e+00   1.607   0.1100
REGIONCAC         -8.540e+00  7.109e+00  -1.201   0.2315
REGIONEU           2.149e+01  1.058e+01   2.032   0.0438 *
REGIONME           8.713e+00  6.147e+00   1.417   0.1584
REGIONNA           2.373e+01  2.585e+01   0.918   0.3599
REGIONOCEANIA      2.373e+00  4.286e+00   0.554   0.5806
REGIONREURO        2.399e+00  1.606e+01   0.149   0.8814
REGIONSA          -1.132e+01  1.816e+01  -0.624   0.5338
FIVEDTH           -9.203e-02  6.592e-03 -13.961  < 2e-16 ***
ADTMORT           -4.563e-02  2.155e-03 -21.172  < 2e-16 ***
BMI                5.671e-02  1.094e-01   0.519   0.6048
GDP                2.973e-05  8.782e-06   3.385   0.0009 ***
ECONOMYDeveloping -3.074e+00  6.005e-01  -5.120 8.85e-07 ***
REGIONASIA:BMI    -2.654e-01  1.690e-01  -1.570   0.1184
REGIONCAC:BMI      3.655e-01  2.663e-01   1.372   0.1719
REGIONEU:BMI      -8.557e-01  4.039e-01  -2.119   0.0357 *
REGIONME:BMI      -3.282e-01  2.287e-01  -1.435   0.1534
REGIONNA:BMI      -8.670e-01  9.282e-01  -0.934   0.3517
REGIONOCEANIA:BMI -1.433e-01  1.616e-01  -0.887   0.3765
REGIONREURO:BMI   -8.781e-02  6.073e-01  -0.145   0.8852
REGIONSA:BMI       4.742e-01  6.794e-01   0.698   0.4862
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.234 on 157 degrees of freedom
Multiple R-squared:  0.9781,    Adjusted R-squared:  0.9752
F-statistic: 333.7 on 21 and 157 DF,  p-value: < 2.2e-16
```

Our R-squared and Adjusted R-square values for this Reduced Model:

R-squared = 0.9781	97.81% of the variation in the response, LIFEEXP, can be explained by the regression model.
Adj R-squared = 0.9752	Looking at the **adjusted R-squared** values for the reduced transformed model with interaction terms and the original transformed model with interaction terms, we can see that its value decreased after removing the 2 predictors, β_{12} HIVINC and β_{23}GDP:ECONOMY_developing. The adjusted R-squared value decreased from .9755 to .**9752**, indicating that we added insignificant predictors to our model since the adjusted R-squared penalizes us for doing so. This makes sense since we chose to keep 7 interaction terms that were insignificant according to the T-test.

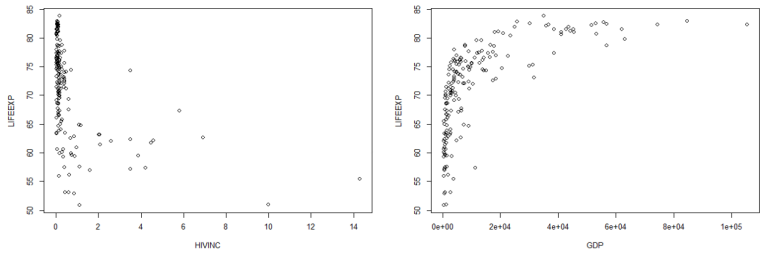
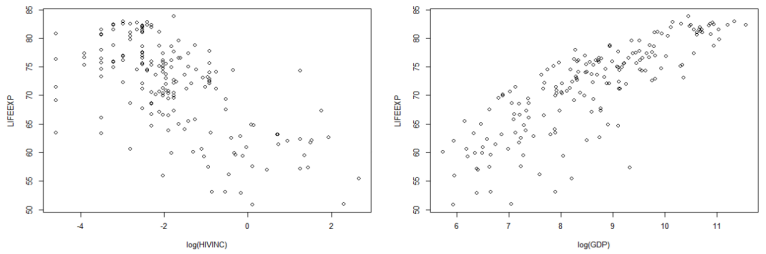
Analyze the LINE Assumptions of the Reduced Model with Interaction Terms ($redM_2$):



Assumption	Graph	Decision	Reasoning
Linearity	Residual vs Fitted	Satisfied	The mean of the points within the slices seems to be roughly equal to 0
Independence	Residuals vs Index	Satisfied	There is no pattern in the spread of the points.
Normality	Histogram of residuals Q-Q plot	Satisfied	The histogram of residuals looks roughly normally distributed Most of the points in the Q-Q plot do not significantly deviate from the line with a slope of 45 degrees
Equal Variance	Residual vs Fitted	Satisfied (better than original transformed model)	The spread of the points within the slices is roughly the same, more so than the original Residual vs Fitted plot

Model 2: Transformation 2: Log(x) for Several Predictors

We also tried a log transformation on some of the predictors in the original, reduced model (redM_1) because we noticed that some of the graphs in the scatterplot matrix consist of points in a narrow band, such as the scatterplot for HIVINC and GDP (to some extent). We did not apply a log transformation to the response variable because when we attempted to do so, the histogram for our logged(LIFEEXP) was even more skewed than the histogram for the original response variable.

Initial Regression Scatterplots: HIVINC GDP*	
Transformed Regression Scatterplots: (Log(Xs))	

*(Regarding the transformation for GDP, we tried both reciprocal and log transformations individually since the initial scatterplot for GDP is curved, but log(GDP) spreads out the points more than the reciprocal transformation, hence we applied a log transformation to GDP in the end.)

Regression Output for the **Transformed Logged Model**:

```
Call:
lm(formula = LIFEEXP ~ REGION + FIVEDTH + ADTMORT + BMI + log(HIVINC) +
    log(GDP) + ECONOMY, data = dat1)

Residuals:
     Min       1Q   Median       3Q      Max
-2.94783 -0.84573  0.05486  0.81686  3.07287

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)       81.821086   2.128927  38.433  < 2e-16 ***
REGIONASIA         0.107978   0.379672   0.284 0.776464
REGIONCAC          1.584094   0.420242   3.769 0.000228 ***
REGIONEU          -1.024778   0.625968  -1.637 0.103525
REGIONME          -0.048251   0.497673  -0.097 0.922882
REGIONNA          -0.053557   0.849319  -0.063 0.949797
REGIONOCEANIA     -1.019303   0.513080  -1.987 0.048628 *
REGIONREURO        0.396989   0.492513   0.806 0.421383
REGIONSA           1.439030   0.474158   3.035 0.002799 **
FIVEDTH           -0.084079   0.006847 -12.279  < 2e-16 ***
ADTMORT           -0.045764   0.002513 -18.214  < 2e-16 ***
BMI               -0.137010   0.067447  -2.031 0.043832 *
log(HIVINC)       -0.047557   0.107479  -0.442 0.658724
log(GDP)           0.640475   0.136088   4.706 5.33e-06 ***
ECONOMYDeveloping -2.979601   0.549020  -5.427 2.03e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.226 on 164 degrees of freedom
Multiple R-squared:  0.9774,    Adjusted R-squared:  0.9755
F-statistic: 506.8 on 14 and 164 DF,  p-value: < 2.2e-16
```

Is the Transformed (Logged) Model Significant?
To answer this question, we performed a **Global F-test** with α = 5%:

Step 1: Hypothesis	Ho: $\beta 1 = \beta 2 = \ldots \beta 14 = 0$ Ha: at least one of the β's $\neq 0$
Step 2: Test Statistic	F(14:164) = 506.8, where k=14 and n=179
Step 3: P-value	P < 2.2e-16 → P = approx. 0
Step 4: Compare p-value and alpha	P-value: approx. 0 < alpha: 0.05 Decision Rule: p-value < alpha → reject Ho **<u>Conclusion:</u>** Since our p-value, 0, is less than alpha, 0.05, we reject our null hypothesis. At least one of the predictors is significant in explaining the response, therefore the **transformed (logged) model is significant.**

Are there any Insignificant Predictors in this Logged Transformed Model?
To answer this question, we conducted **T-tests for each of the predictors.** As we had 14 predictors, the table below is a summary of our T-test results:

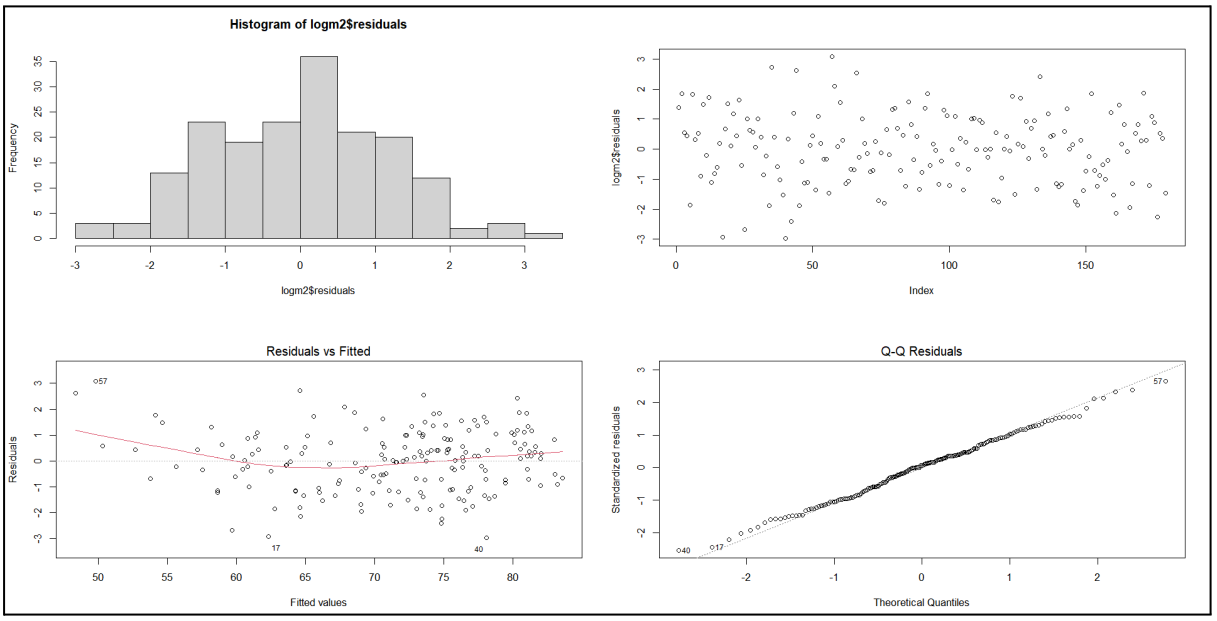
Predictor	**T-test (α = 5%)**
Hypotheses for all the T-tests:	Hypothesis: - Ho: $\beta_i = 0$ - Ho: $\beta_i \neq 0$
β_2 REGION_CAC β_6 REGION_OCEANIA β_8 REGION_SA β_9FIVEDTH β_{10}ADTMORT β_{11}BM β_{13}Log(GDP) β_{14}Economy_Developing	Decision Rule: Since all of these predictor's p-value < alpha 0.05, we reject Ho **Conclusion**: These predictors (to the left) are each significant predictors in explaining the response after adjusting for the other predictors
β_1REGION_ASIA *β_3REGION_EU* *β_4REGION_ME* *β_5REGION_NA* *β_7REGIONREURO*	**While some of the Region dummy variables were deemed insignificant predictors according to the T-test, we decided to keep them in our model because three of them were significant.*
β_{13}Log(HIVINC)	Decision Rule: Since β_{13}Log(HIVINC)'s p-value > alpha 0.05, we fail to reject Ho **Conclusion**: β_{13}Log(HIVINC) is not a significant predictor in explaining the response after adjusting for the other predictors

Since it is a just single variable, β_{13}Log(HIVINC), we can **remove it from the model** without conducting a partial F-test since the T-test has concluded that β_{13}Log(HIVINC) is an insignificant predictor and thus can be removed from the model.

Reduced Model after a Log Transformation (redM$_3$) is:

LIFEEXP: β_0+ β_1REGION_ASIA + β_2 REGION_CAC - β_3REGION_EU + β_4REGION_ME + β_5REGION_NA - β_6REGION_OCEANIA + β_7REGION_REURO + β_8REGION_SA - β_9FIVEDTH - β_{10} ADTMORT - β_{11} BMI + β_{12} Log(GDP) - β_{13}ECONOMY_ developing + e

Graphs to Analyze the LINE Assumptions of the Transformed (Logged), Reduced Model:



Assumption	Graph	Decision	Reasoning
Linearity	Residual vs Fitted	Satisfied (better than the original reduced model)	The mean of the points within the slices seems to be roughly equal to 0, more than the original reduced model
Independence	Residuals vs Index	Satisfied	There is no pattern in the spread of the points.
Normality	Histogram of residuals Q-Q plot	Satisfied (a little worse than the original reduced model)	The histogram of residuals looks roughly normally distributed, a little less so than the original reduced histogram Most of the points in the Q-Q plot do not significantly deviate from the line with a slope of 45 degrees, but a little less so than the original reduced Q-Q plot

Equal Variance	Residual vs Fitted	Satisfied (better than the original reduced model)	The spread of the points within the slices is roughly the same, more so than the original reduced Residual vs Fitted plot

Regression Output for the **Log Transformed, Reduced Model ($redM_3$)** created in R Studio:

```
Call:
lm(formula = LIFEEXP ~ REGION + FIVEDTH + ADTMORT + BMI + log(GDP) +
    ECONOMY, data = dat1)

Residuals:
     Min       1Q   Median       3Q      Max
-2.97731 -0.86472  0.08033  0.82062  3.08794

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)       82.110586   2.020956  40.630  < 2e-16 ***
REGIONASIA         0.164242   0.356873   0.460 0.645960
REGIONCAC          1.607501   0.415882   3.865 0.000159 ***
REGIONEU          -0.964291   0.609369  -1.582 0.115464
REGIONME           0.007961   0.480011   0.017 0.986788
REGIONNA          -0.015894   0.842981  -0.019 0.984979
REGIONOCEANIA     -0.979937   0.504074  -1.944 0.053593 .
REGIONREURO        0.461947   0.468978   0.985 0.326063
REGIONSA           1.466180   0.469024   3.126 0.002094 **
FIVEDTH           -0.083546   0.006724 -12.425  < 2e-16 ***
ADTMORT           -0.046379   0.002089 -22.203  < 2e-16 ***
BMI               -0.137942   0.067250  -2.051 0.041830 *
log(GDP)           0.626524   0.132062   4.744 4.51e-06 ***
ECONOMYDeveloping -2.996524   0.546350  -5.485 1.53e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.223 on 165 degrees of freedom
Multiple R-squared:  0.9774,    Adjusted R-squared:  0.9756
F-statistic: 548.5 on 13 and 165 DF,  p-value: < 2.2e-16
```

Our R-squared and Adjusted R-Square values for the Reduced, Logged Model:

R-squared = 0.9774	97.74% of the variation in the response, LIFEEXP, can be explained by the regression model.
Adj R-squared = 0.9756	Our initial (reduced) model's adjusted R-square value was .9751. After transforming the model, it increased to .9756, which indicates that taking out the log(HIVINC) was a good move since the adjusted R-squared value increased after doing so. (Since if an insignificant predictor was kept in the model, the adj R-squared would 'penalize' us and decrease, so the fact that adj R-squared increased after removing log(HIVINC) indicates that we did remove an insignificant predictor.)

Model 3: Variable Selection Technique: Backward Elimination

In addition to our two different transformed models, we also wanted to use a variable selection technique, backward elimination, to see which predictors would a computer-generated program deem insignificant and worthy of elimination from the model.

The model we used for Backward Elimination was our **initial model**:

LIFEEXP = β_0+ β_1REGION_ASIA + β_2REGION_CAC - β_3REGION_EU + β_4REGION_ME + β_4REGION_NA - β_6REGION_OCEANIA + β_7REGION_REURO + β_8REGION_SA - β_9INFDTH - β_{10}FIVEDTH - β_{11}ADTMORT - β_{12}ALCOHOL - β_{13}HBV + β_{14}MMR - β_{15}BMI - β_{16}POLIO + β_{17}DIPH + β_{18}HIVINC+ β_{19}GDP - β_{20}POP + β_{21}EDU - β_{22}ECONOMY_ developing + e

Backward Elimination Summary Output from R:

Stepwise Summary

Step	Variable	AIC	SBC	SBIC	R2	Adj. R2
0	Full Model	609.636	686.133	92.330	0.97788	0.97476
1	POLIO	607.642	680.952	90.040	0.97788	0.97492
2	ALCOHOL	605.712	675.835	87.804	0.97787	0.97507
3	DIPH	603.837	670.772	85.614	0.97785	0.97521
4	POP	602.011	665.759	83.470	0.97783	0.97534
5	INFDTH	600.592	661.153	81.682	0.97776	0.97541

Final Model Output

Model Summary

R	0.989	RMSE	1.165
R-Squared	0.978	MSE	1.508
Adj. R-Squared	0.975	Coef. Var	1.719
Pred R-Squared	0.972	AIC	600.592
MAE	0.926	SBC	661.153

ANOVA

	Sum of Squares	DF	Mean Square	F	Sig.
Regression	10676.454	17	628.027	416.341	0.0000
Residual	242.860	161	1.508		
Total	10919.314	178			

Parameter Estimates

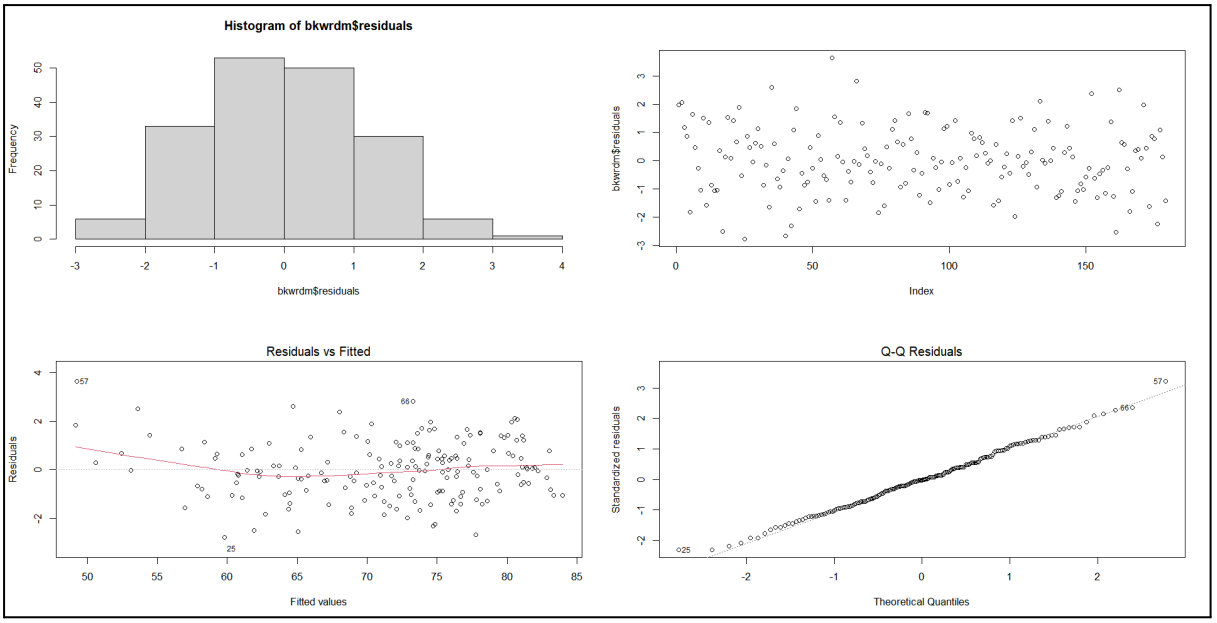
model	Beta	Std. Error	Std. Beta	t	Sig	lower	upper
(Intercept)	87.261	2.033		42.917	0.000	83.246	91.277
REGIONASIA	0.122	0.377	0.006	0.323	0.747	-0.623	0.867
REGIONCAC	1.790	0.438	0.071	4.089	0.000	0.925	2.654
REGIONEU	-0.734	0.645	-0.034	-1.138	0.257	-2.007	0.539
REGIONME	0.119	0.497	0.004	0.239	0.811	-0.862	1.100
REGIONNA	-0.049	0.881	-0.001	-0.055	0.956	-1.789	1.692
REGIONOCEANIA	-0.973	0.527	-0.030	-1.845	0.067	-2.014	0.068
REGIONREURO	0.163	0.523	0.006	0.311	0.756	-0.871	1.196
REGIONSA	1.844	0.489	0.059	3.767	0.000	0.877	2.810
FIVEDTH	-0.086	0.009	-0.352	-9.921	0.000	-0.103	-0.068
ADTMORT	-0.050	0.003	-0.576	-17.362	0.000	-0.056	-0.044
HBV	-0.011	0.009	-0.020	-1.256	0.211	-0.029	0.006
MMR	0.011	0.008	0.023	1.444	0.151	-0.004	0.027
BMI	-0.157	0.074	-0.044	-2.133	0.034	-0.303	-0.012
HIVINC	0.190	0.085	0.039	2.227	0.027	0.022	0.359
GDP	0.000	0.000	0.058	2.978	0.003	0.000	0.000
EDUC	0.085	0.068	0.034	1.255	0.211	-0.049	0.220
ECONOMYDeveloping	-2.681	0.606	-0.139	-4.422	0.000	-3.879	-1.484

Backward Elimination Technique fits the full model with all k predictors and then looks for the most insignificant predictor (whose corresponding T-test has the highest p-value.) It compares this predictor's p-value to $alpha_{remove}$ (default = 0.3) and eliminates the predictor if p > $alpha_{remove}$. The computer then fits the model with the remaining k-1 predictors and continues the process until there are no more predictors that can be eliminated from the model. According to this technique, **five predictors**: POLIO, ALCOHOL, DIPH, POP, and INFDTH should be eliminated from our initial model.

Our reduced model **($redM_4$)**after this variable selection technique:

LIFEEXP = β_0+ β_1REGION_ASIA + β_2REGION_CAC - β_3REGION_EU + β_4REGION_ME + β_5REGION_NA - β_6REGION_OCEANIA + β_7REGION_REURO + β_8REGION_SA - β_9FIVEDTH - β_{10}ADTMORT - β_{11}HBV + β_{12}MMR - β_{13}BMI + β_{14}HIVINC+ β_{15}GDP + β_{16}EDU - β_{17}ECONOMY_ developing + e

Analyze the LINE Assumptions of the Reduced Model from Backward Elimination:



Assumption	Graph	Decision	Reasoning
Linearity	Residual vs Fitted	Satisfied	The mean of the points within the slices seem to be roughly equal to 0, very similar to the original chart
Independence	Residuals vs Index	Satisfied	There is no pattern in the spread of the points.
Normality	Histogram of residuals Q-Q plot	Satisfied	The histogram of residuals looks roughly normally distributed Most of the points in the Q-Q plot do not significantly deviate from the line with a slope of 45 degrees
Equal Variance	Residual vs Fitted	Satisfied	The spread of the points within the slices is roughly the same.

We then conducted a **Partial F-test** to ensure that we could use this reduced model over the original model:

Models	**Model$_{FULL}$**: β_0+ β_1REGION_ASIA + β_2 REGION_CAC - β_3REGION_EU + β_4REGION_ME + β_5REGION_NA - β_6REGION_OCEANIA + β_7REGION_REURO + β_8REGION_SA - β_9 INFDTH - β_{10} FIVEDTH - β_{11} ADTMORT - β_{12} ALCOHOL - β_{13} HBV + β_{14} MMR - β_{15} BMI - β_{16} POLIO + β_{17}DIPH + β_{18} HIVINC+ β_{19} GDP - β_{20}POP + β_{21}EDUC - β_{22}ECONOMY_ developing + e **Model$_{REDUCED}$**: β_0+ β_1REGION_ASIA + β_2REGION_CAC - β_3REGION_EU + β_4REGION_ME + β_5REGION_NA - β_6REGION_OCEANIA + β_7REGION_REURO + β_8REGION_SA - β_{10}FIVEDTH - β_{11}ADTMORT - β_{13}HBV + β_{14}MMR - β_{15}BMI + β_{18}HIVINC+ β_{19}GDP + β_{21}EDU - β_{22}ECONOMY_ developing + e
Step 1: Hypothesis	Ho: $\beta 9 = \beta 12 = \beta 16 = \beta 17 = \beta 20 = 0$ Ha: at least one of the β's $\neq 0$
Step 2: Test Statistic	F (5:156) = 0.1672 (from ANOVA table in R)
Step 3: P-value	P-value = 0.9743 (from ANOVA table in R)
Step 4: Compare p-value and alpha	P-value: 0.9743 > alpha: 0.05 Decision Rule: p-value > alpha → fail to reject Ho (hence use reduced model) **Conclusion**: Since our p-value for this model was greater than alpha, we failed to reject Ho, hence we can eliminate all 5 predictors from the model at the same time. **We can use the Reduced Model that resulted after the Backward Elimination Technique, redM$_4$**
ANOVA TABLE	Model 1: LIFEEXP ~ REGION + FIVEDTH + ADTMORT + HBV + MMR + BMI + HIVINC + GDP + EDUC + ECONOMY Model 2: LIFEEXP ~ REGION + INFDTH + FIVEDTH + ADTMORT + ALCOHOL + HBV + MMR + BMI + POLIO + DIPH + HIVINC + GDP + POP + EDUC + ECONOMY Res.Df RSS Df Sum of Sq F Pr(>F) 1 161 242.86 2 156 241.56 5 1.2946 0.1672 0.9743

Regression Output for the **Reduced Model after Elimination of the 5 Predictors ($redM_4$):**

```
Call:
lm(formula = LIFEEXP ~ REGION + FIVEDTH + ADTMORT + HBV + MMR +
    BMI + HIVINC + GDP + EDUC + ECONOMY, data = dat1)

Residuals:
    Min      1Q  Median      3Q     Max
-2.7829 -0.8446 -0.0278  0.7748  3.6486

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)        8.726e+01  2.033e+00  42.917  < 2e-16 ***
REGIONASIA         1.217e-01  3.772e-01   0.323 0.747312
REGIONCAC          1.790e+00  4.377e-01   4.089 6.82e-05 ***
REGIONEU          -7.339e-01  6.447e-01  -1.138 0.256697
REGIONME           1.188e-01  4.968e-01   0.239 0.811359
REGIONNA          -4.861e-02  8.813e-01  -0.055 0.956083
REGIONOCEANIA     -9.730e-01  5.273e-01  -1.845 0.066833 .
REGIONREURO        1.626e-01  5.232e-01   0.311 0.756407
REGIONSA           1.844e+00  4.894e-01   3.767 0.000231 ***
FIVEDTH           -8.552e-02  8.620e-03  -9.921  < 2e-16 ***
ADTMORT           -5.014e-02  2.888e-03 -17.362  < 2e-16 ***
HBV               -1.124e-02  8.944e-03  -1.256 0.210789
MMR                1.132e-02  7.841e-03   1.444 0.150711
BMI               -1.572e-01  7.370e-02  -2.133 0.034458 *
HIVINC             1.903e-01  8.543e-02   2.227 0.027323 *
GDP                2.583e-05  8.672e-06   2.978 0.003345 **
EDUC               8.541e-02  6.805e-02   1.255 0.211215
ECONOMYDeveloping -2.681e+00  6.064e-01  -4.422 1.79e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.228 on 161 degrees of freedom
Multiple R-squared:  0.9778,    Adjusted R-squared:  0.9754
F-statistic: 416.3 on 17 and 161 DF,  p-value: < 2.2e-16
```

Our R-squared and Adjusted R-square values for this Reduced Model:

R-squared = 0.9778	97.78 % of the variation in the response, LIFEEXP, can be explained by the regression model.
Adj R-squared = 0.9754	Our initial model's adjusted R-square value was .9748. After transforming the model, it increased to .9754, which indicates that taking out the 5 predictors was a good move since the adjusted R-squared value increased by .0006 after doing so.

Conclusion of Phase 1

Of the three models we created, we compared the R-squared, Adjusted R-squared, and the Standard Error for each to determine our final model that best fits the data.

Model	**R-squared**	**Adj R-Squared**	**Std. Error**
Initial Reduced Model ($redM_1$)	97.71%	.9751	1.235
Interaction Term Model ($redM_2$)	**97.81%**	.9752	1.234
Log(x) Model ($redM_3$)	97.74%	**.9756**	**1.223**
Backward Elimination Model ($redM_4$)	97.78%	.9754	1.228

Assessing the Model

After making improvements to the already satisfied LINE assumptions as best as we could, we are leaning towards picking **redM$_3$, which was our reduced log(x) model**. However, we wanted to check all three of our modified models (redM$_2$,redM$_3$,redM$_4$) for overfitting and omitted variable bias to ensure that they are all good models to choose from. Afterward, we will decide on a model and identify multicollinearity, possible outliers, and influential points in that model. If we discover any, we will address them to make our chosen model even better than it is.

Overfitting (Akaike Information Criteria)

Overfitting occurs when we keep unimportant variables in the model, which will lead to inaccurate predictions when given a new data set even if the model fits the current data really well. To check for overfitting, we used the **AIC(model)** function in R. Here are our results for each of our three models for comparison sake:

Models	**AIC Value**
Interaction Term Model (redM$_2$)	605.917
Log(x) Model (redM$_3$)	**595.603**
Backward Elimination Model (redM$_4$)	600.5924

The smaller the AIC value is, the better and more fit the model is. As we can see in the table, the model that has the smallest AIC value is the one that we prefer best at the moment, our **log(x) transformed reduced model (redM$_3$)**.

Omitted Variable Bias (Mallow's Cp)

Omitted Variable Bias occurs under two conditions: when the omitted variable should appear in the model but does not (when the regression coefficient of the omitted variable must =/= 0) and when the omitted variable highly correlates with other variables in the model. We used Mallow's Cp to identify if a model has omitted variable bias.

The Criteria for Cp is:
If Cp <= p, then there is no omitted variable bias.
If Cp > p, then there is omitted variable bias.

We tested all three of our potential final models for omitted variable bias.
We treated the initial reduced model as our Full Model since it is the most polished version.
In R, we used the function **ols_mallows_cp(potential model, initial reduced model)**

Models	Cp	**P** (= K+1)	Omitted Variable Bias?
Interaction Term Model **(redM$_2$)**	21.76156	22 (= 21+1)	Cp < 22 → No Omitted Variable Bias
Log(x) Model **(redM$_3$)**	10.82107	14 (=13+1)	Cp < 14 → No Omitted Variable Bias
Backward Elimination Model **(redM$_4$)**	16.12216	18 (=17+1)	Cp < 18 → No Omitted Variable Bias

None of the models we created suffer from omitted variable bias, which is a sign that we are on the right track.

Chosen Model:

After reviewing the various tests (global F-test, T-test, partial F-test), R-squared values, adjusted R-squared values, standard errors, AIC, and Mallow's Cp for each of the 3 potential models, we have decided on a final model:

Log(x) transformed reduced model (**redM$_3$**)

LIFEEXP: β_0+ β_1REGION_ASIA + β_2 REGION_CAC - β_3REGION_EU + β_4REGION_ME + β_5REGION_NA - β_6REGION_OCEANIA + β_7REGION_REURO + β_8REGION_SA - β_9FIVEDTH - β_{10} ADTMORT - β_{11} BMI + β_{12} Log(GDP) - β_{13}ECONOMY_ developing + e

While this model had the third highest R-squared value (97.81%), it had the highest adjusted R-squared value (.9752), lowest standard error value (1.223), lowest AIC value (595.603), and does not contain omitted variable bias. While the other models also had values that were close to the Log(x) transformed reduced model, it is overall the best model of the three.

Multicollinearity

We used **VIF (Variance Inflation Factor)** on our **potential final model,** (**redM$_3$**) to determine if it contained multicollinearity:

```
> vif(logm2)
               GVIF Df GVIF^(1/(2*Df))
REGION    16.073750  8        1.189549
FIVEDTH    5.580721  1        2.362355
ADTMORT    4.198463  1        2.049015
BMI        2.583317  1        1.607270
log(GDP)   4.002502  1        2.000625
ECONOMY    5.853335  1        2.419367
```

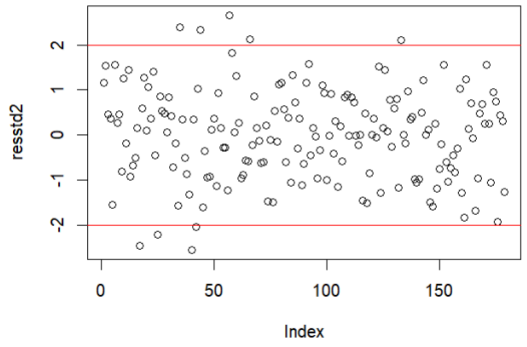
Our results came out as a general VIF due to our model having categorical variables. As a result, we must square the GVIF^(1/(2*Df)) values first before comparing the values against our criterion for multicollinearity (>10 VIF):

Predictor	**(GVIF^(1/(2*Df)))^2 → "VIF"**
REGION	1.415026823
FIVEDTH	5.580721146
ADTMORT	4.19846247
BMI	2.583316853
log(GDP)	4.002500391
ECONOMY	5.853336681

Since none of our predictor's squared GVIF^(1/(2*Df)) values are greater than 10, we do not have multicollinearity in this potential final model. That indicates that no two predictors are strongly linearly correlated with each other, which decreases our chances of having unreliable estimates in this model.

Outliers

Standardized Residual vs Index Plot



If $|SR_i| > 2$, obs_i is considered an outlier.

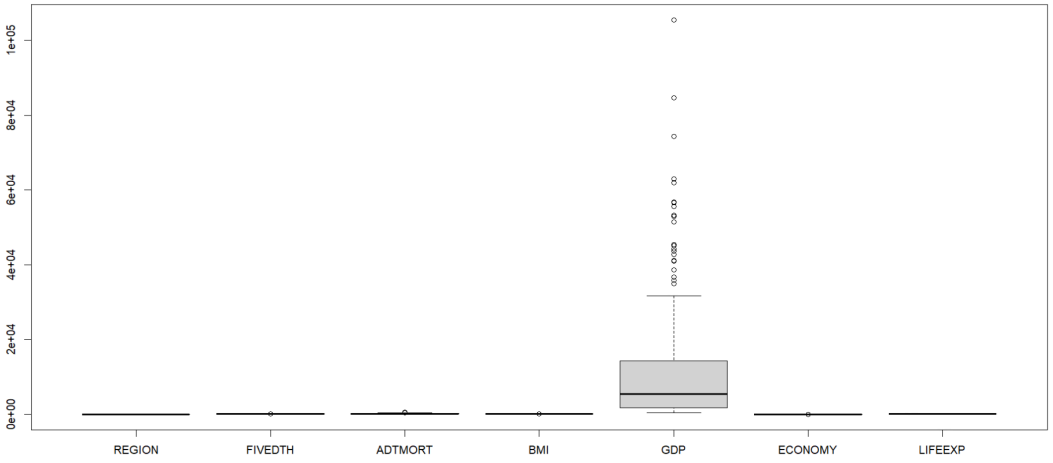
There are 8 points that are outside the red lines that show SR=-2 and SR=2, therefore **our model has 8 outliers.**

```
resstd2 <-rstandard(logm2)
plot(resstd2)
abline(a=2,b=0, col="red")
abline(a=-2,b=0,col="red")
```

There are several ways to investigate why we have these outliers:

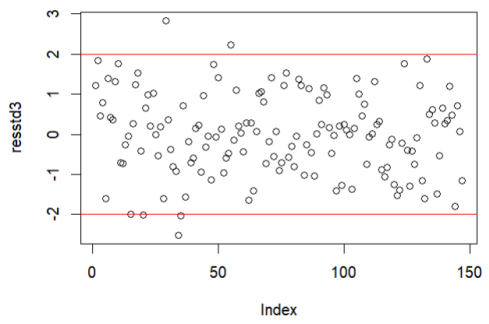
1) Data entry mistake → This seems unlikely since this dataset came from the World's Health Organization and World Bank.
2) Missing predictor → Through Mallow's Cp, we already established that this model is not omitting any important predictors (no omitted variable bias)
3) One or more assumptions are not met → All assumptions are met
4) The observation differs significantly from majority of the observations:
    a) Remove these observations and refit the model
    b) Use the new model *if* it differs significantly from the initial model
    c) Otherwise, use the model with all observations included since getting rid of some of the points may drastically change the data

To get rid of these outliers, we first created a boxplot using boxplot() in R with a subset of our data that only included the columns in our final model (**redM$_3$**) since that is what our Standardized Residuals vs Index Plot was based on:




```
> boxplot(dat2)$out
 [1]    128.7000    140.2000    126.4000    134.3000    123.3000    434.8210    513.4755    411.0745     32.1000
[10]     32.1000  74356.0000  35808.0000  38631.0000  51545.0000  43596.0000  45193.0000  38663.0000  42802.0000
[19]  41008.0000  34961.0000  44196.0000  62012.0000  84776.0000  63039.0000  45405.0000  41103.0000  53255.0000
[28]  55647.0000  56707.0000 105462.0000  56763.0000  52952.0000  36653.0000      1.0000      1.0000      1.0000
[37]      1.0000      1.0000      1.0000      1.0000      1.0000      1.0000      1.0000      1.0000      1.0000
[46]      1.0000      1.0000      1.0000      1.0000      1.0000      1.0000      1.0000      1.0000      1.0000
[55]      1.0000      1.0000      1.0000      1.0000      1.0000      1.0000      1.0000      1.0000      1.0000
[64]      1.0000      1.0000      1.0000      1.0000      1.0000      1.0000      1.0000
```

After removing the outliers for our numerical predictors FIVEDTH, ADTMORT, BMI, GDP, our Standardized Residual vs Index Plot looks like this:




```
logm2v2 <-
lm(LIFEEXP~REGION+FIVEDTH+ADTMORT+
BMI+log(GDP)+ECONOMY,data=dat2v2)
> resstd3<-rstandard(logm2v2)
> plot(resstd3)
> abline(a=2,b=0, col="red")
> abline(a=-2,b=0, col="red")
```

If $|SR_i| > 2$, obs_i is considered an outlier.

There are 3 points that are outside the red lines that show SR=-2 and SR=2, so we still have 3 outliers, but it is better than when the model had 8 outliers.

*To remove these outliers, I first saved the outliers for each of the predictors in a vector like this:

```
outlier1 <- boxplot(dat2$FIVEDTH,
plot=FALSE)$out
```

which I then excluded from our data set like this:

```
dat2v2<-dat2v2[-which(dat2v2$FIVEDTH %in%
outlier1),]
```

I did this for all numerical predictors FIVEDTH, ADTMORT, DMI, and GDP and then re-graphed the Standardized Residual vs Index plot with the updated model and dataset

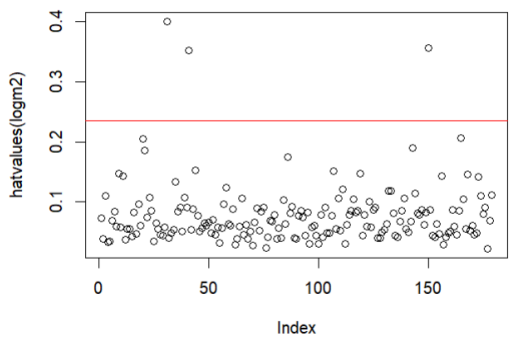
In the table below, we compared the Log(x) model before and after removing the data points:

Model	**R-squared**	**Adj R-Squared**	**Std. Error**
Log(x) Model ($redM_3$)	**97.74%**	**.9756**	1.223
Log(x) Model AFTER removing outliers ($redM_5$)	96.82%	.9653	**1.208**

After a discussion, our team decided to keep the original 8 outliers in the model and use the **initial Log(x) model ($redM_3$)** because the values in the table above were not drastically different, with a less than 0.02 change in all values. Additionally, the $redM_3$ model had a higher R-squared and Adj R-squared value than the model where we removed the outliers.

We continued to use the initial Log(x) model and the initial dataset to test for high leverage points and influential points:

Leverage Points



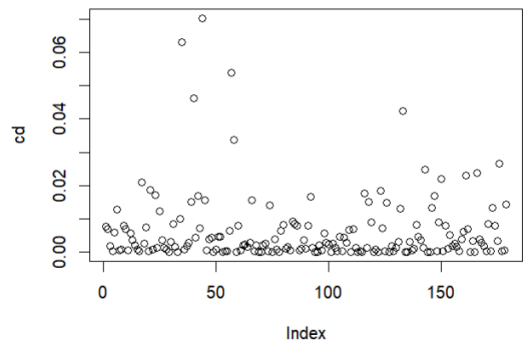
If $h_i > (3*(k+1))/n$, then obs_i has a high leverage

Calculation: (3*(14+1))/179 = 0.2346

There are 3 points above the 0.2346 line, therefore **our model has 3 high leverage points.**

```
plot(hatvalues(logm2))
abline(a=.2346,b=0,col="red")
```

Influential Points (Cook's Distance)



A point is considered a potential influential point if $Cd_i > 0.5$ (on the Y-axis) and an influential point if $Cd_i > 1$.

Since all the points in this plot are less than 0.5, **there are no influential points in this model (all points are 0.07 or below).**

```
cd<-cooks.distance(logm2)
plot(cd)
abline(a=1/2,b=0)
```

Model Selection

To reiterate, our Final Model is **redM$_3$**:

LIFEEXP = 82.11 + 0.164REGION_ASIA + 1.608REGION_CAC - 0.964REGION_EU + 0.008REGION_ME - 0.016REGION_NA - 0.98REGION_OC + 0.462REGIONREURO + 1.466REGION_SA - 0.084FIVEDTH - 0.046ADTMORT - 0.138BMI + 0.627log(GDP) - 2.997ECONOMY_ developing

The table below restates the **significant numerical predictors** used in this final model:

Predictors	**Description** (per country)	**Units**
FIVEDTH	the number of children (< 5 years) deaths	deaths per 1000 population
ADTMORT	the number of adult deaths	deaths per 1000 population
BMI	the average BMI for that country	kg/m^2
GDP	GDP per capita of that country	United States Dollars

The table below restates the **significant categorical predictors** used in this final model:

Predictors	**Description**	**Baseline (left out of model)**	**Other Indicators**
ECONOMY	the economic conditions of that country	Developed	Developing
REGION*	the region of the country was distributed into	AFR (Africa)	-ASIA -ME (Middle East) -EU (European Nation) -REURO (Rest of Europe) -SA (South America) -NA (North America) -OCEANIA -CAC (Central America and the Caribbean)

Interpreting the predictors' effect on our response variable, Life Expectancy:

- **LIFEEXP** → The average life expectancy of a country of both genders (since birth) in years
- **REGION_ASIA** → Holding all other predictors fixed, the average life expectancy for countries in Asia is expected to be 0.164 years higher than the average life expectancy for countries in Africa.
- **REGION_CAC** → Holding all other predictors fixed, the average life expectancy for countries in Central America and the Caribbean is expected to be 1.608 years higher than the average life expectancy for countries in Africa.
- **REGION_EU** → Holding all other predictors fixed, the average life expectancy for countries in Europe is expected to be 0.964 years lower than the average life expectancy for countries in Africa.
- **REGION_ME** → Holding all other predictors fixed, the average life expectancy for countries in the Middle East is expected to be 0.008 years lower than the average life expectancy for countries in Africa.
- **REGION_NA** → Holding all other predictors fixed, the average life expectancy for countries in North America is expected to be 0.016 years lower than the average life expectancy for countries in Africa.
- **REGION_OC** → Holding all other predictors fixed, the average life expectancy for countries in Oceania is expected to be 0.980 years lower than the average life expectancy for countries in Africa.
- **REGION_REURO** → Holding all other predictors fixed, the average life expectancy for countries in the rest of Europe is expected to be 0.462 years higher than the average life expectancy for countries in Africa.
- **REGION_SA** → Holding all other predictors fixed, the average life expectancy for countries in South America is expected to be 1.466 years higher than the average life expectancy for countries in Africa.
- **FIVEDTH** → As the number of deaths for children under the age of 5 increases by 1 (per 1,000 population), while holding the other predictors fixed, the life expectancy, on average, decreases by 0.084 years.
- **ADTMORT** → As the number of adult deaths increases by 1 (per 1,000 population), while holding the other predictors fixed, the life expectancy, on average, decreases by 0.046 years.
- **BMI** → As BMI increases by 1 Kg/m^2, while holding the other predictors fixed, the life expectancy, on average, decreases by 0.138 years.
- **log(GDP)** → A 1% increase in GDP is associated with an average increase in life expectancy (LIFEEXP) of (0.627/100) years while holding the other predictors fixed.
- **ECONOMY_developing** → Holding all other predictors fixed, the average life expectancy for countries with developing economies is expected to be 2.997 years lower than the average life expectancy for countries with developed economies.

Estimating Expected Values of Y (general level):

We found the expected value of Y, life expectancy, for three general scenarios. We made sure that the simulated values we chose fell in the range of the dataset (to the extent of looking at each region's observations separately) to avoid extrapolation.

LIFEEXP	REGION	FIVEDTH	ADTMORT	BMI	GDP	ECONOMY
74.49 years	ASIA	25	100	23	$3,500	Developing

Interpretation: On average, the life expectancy of <u>all</u> developing countries in the Asia region with 25 children (under 5 years old) deaths and 100 adult deaths per 1,000 population, an average BMI of 23 kg/m^2, and a GDP per capita of $3,500 is 74.49 years
`> obs<-data.frame(REGION="ASIA",FIVEDTH=25,ADTMORT=100,BMI=23,GDP=3500,ECONOMY="Developing")`
`> predict(logm2,obs)`
`1`
`74.49187`						

75.69 years	CAC	16	120	26	$6,100	Developing

Interpretation: On average, the life expectancy of <u>all</u> developing countries in the Central America and Caribbean region with 16 children (under 5 years old) deaths and 120 adult deaths per 1,000 population, an average BMI of 26 kg/m^2 and a GDP per capita of $6,100 is 75.69 years
`> obs2<-data.frame(REGION="CAC",FIVEDTH=16,ADTMORT=120,BMI=26,GDP=6100,ECONOMY="Developing")`
`> predict(logm2,obs2)`
`1`
`75.6937`						

80.12 years	EU	5	75	26	$30,000	Developed

Interpretation: On average, the life expectancy of <u>all</u> developed countries in the European Union region with 5 children (under 5 years old) deaths and 75 adult deaths per 1,000 population, an average BMI of 26 kg/m^2, and a GDP per capita of $30,000 is 80.12 years
`> obs3<-data.frame(REGION="EU",FIVEDTH=5,ADTMORT=75,BMI=26,GDP=30000,ECONOMY="Developed")`
`> predict(logm2,obs3)`
`1`
`80.12247`						

The expected values of life expectancy (y) we computed (for at birth) make sense in the context of our problem since they are realistic ages for how long humans live. For example, according to the United Nations, the life expectancy for people living in Central America and the Caribbean in 2021 was 75.24 years (since birth), which is pretty close to our expected value for all developing countries in that same region (where the other predictors are known).

Predicting Individual Values of Y at Particular Predictor Values:

We found the individual value of Y, life expectancy, for three individual scenarios. Like before, we made sure that the simulated values we chose fell in the range of the dataset to avoid extrapolation:

LIFEEXP	REGION	FIVEDTH	ADTMORT	BMI	GDP	ECONOMY
77.74 years	OCEANIA	21.5	89.1875	23.9	$11,106	Developed

Interpretation: The average life expectancy of a developed countries in the Oceania region with 21.5 children (under 5 years old) deaths and 89.1875 adult deaths per 1,000 population, an average BMI of 23.9 kg/m^2, and a GDP per capita of $11,106 is 77.74 years

```
> obs4<-data.frame(REGION="OCEANIA",FIVEDTH=21.5,ADTMORT=89.1875,BMI=23.9,GDP=11106,ECONOMY="Developed")
> predict(logm2,obs4)
       1
77.73741
```

80.76 years	EU	4	64	27	$40,000	Developed

Interpretation: The average life expectancy of a developed country in the European Union region with 4 children (under 5 years old) deaths and 64 adult deaths per 1,000 population, an average BMI of 27 kg/m^2, and a GDP per capita of $40,000 is 80.76 years

```
> obs5<-data.frame(REGION="EU",FIVEDTH=4,ADTMORT=64,BMI=27,GDP=40000,ECONOMY="Developed")
> predict(logm2,obs5)
       1
80.75848
```

72.96 years	ME	15	125	27.5	$1750	Developing

Interpretation: The average life expectancy of a developing country in the Middle East region with 15 children (under 5-years-old) deaths and 125 adult deaths per 1,000 population, an average BMI of 27.5 kg/m^2, and a GDP per capita of $1,750 is 72.96 years.

```
> obs6<-data.frame(REGION="ME",FIVEDTH=15,ADTMORT=125,BMI=27.5,GDP=1750,ECONOMY="Developing")
> predict(logm2,obs6)
       1
72.95657
```

Conclusion

We utilized transformations and tests (global F-test, T-test, partial F-test) as well as variable selection techniques to create three potential models we could use over the original. We compared their R-squared, Adjusted R-squared, Standard Error, AIC value, and Cp to determine which model best fits our data. We chose the reduced log(x) transformed model, **redM$_3$**, as our final model. We then checked for any outliers, high-leverage points, and influential points, and while we did find some outliers and high-leverage points, we decided to keep our redM$_3$ the way it was because removing the outliers did not drastically change the model itself.

With this final model, we interpreted each of the predictors, estimated expected values of Y, and predicted individual values of Y at particular values while making sure we did not extrapolate the dataset. Now, we will answer the research questions we proposed in the beginning:

Which factors influence life expectancy the most?

Predictors with higher coefficients have a greater effect on predicting life expectancy. In our model, the following predictors are the only ones to have a coefficient that is greater than the absolute value of 1: **ECONOMY_developing (-2.997), REGION_CAC (1.608), and REGION_SA (1.466).** This means that whether a country is considered to be economically developed or developing, and whether a country is from Central America and the Caribbean (REGION_CAC) or from the South American region (REGION_SA), has a great impact on the average life expectancy (in years) of that country.

Which region has the greatest effect on increasing life expectancy?

Holding all other predictors fixed, the region that **increases life expectancy the greatest is the Central America and Caribbean (REGION_CAC)** as it is the region with the highest positive coefficient in our model. Holding all other predictors fixed, the average life expectancy for countries in Central America and the Caribbean is expected to be 1.608 years higher than the average life expectancy for countries in Africa. The other regions either have a negative effect on life expectancy compared to Africa (negative coefficients) or they have a smaller positive coefficient value than that of REGION_CAC.